Paul Matt
Brian Holtman
Erni Muja
Xi Li

Student Performance on Exams

Introduction to the data set:

Students Performance is a simulated dataset containing the math, reading and writing scores of 1000 students in high school in the US. The students are categorized into gender, race, parental level of education, lunch type and whether the student completed test preparation courses.

Analysis Introduction:

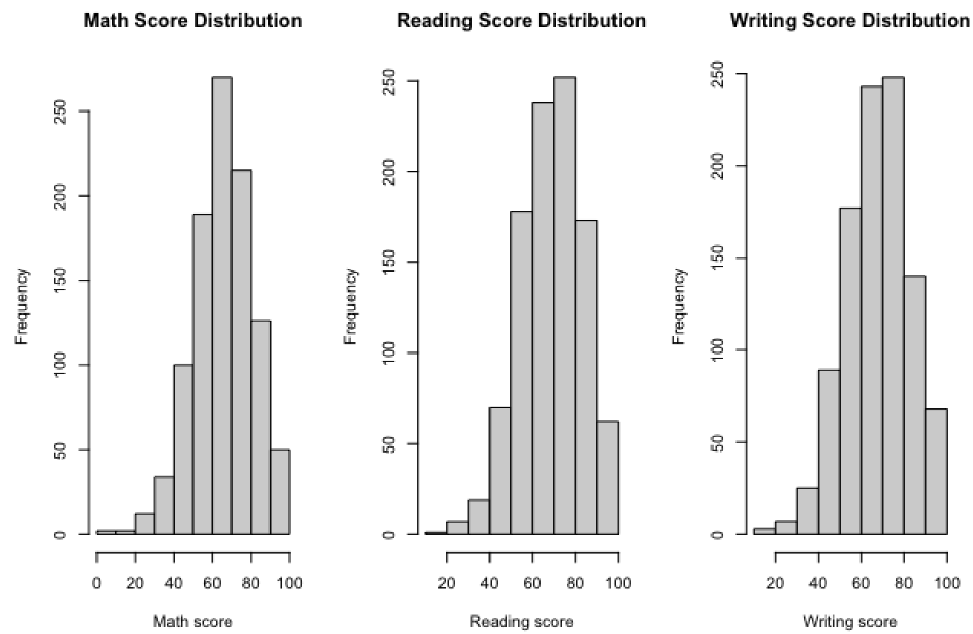
The purpose of our analysis is to predict math scores based on reading and writing scores as well as gender, race, parental level of education, lunch type and whether the student completed test preparation courses. In addition this analysis also seeks to predict gender based on math, writing and reading scores along with race, parental level of education, lunch type and whether the student completed test preparation courses.

Can we accurately predict a student's math score and Gender based on the other variables? Which predictor is statistically significant? Will prediction accuracy and model interpretability improve by removing certain variables?

Summary Statistics of quantitative data:

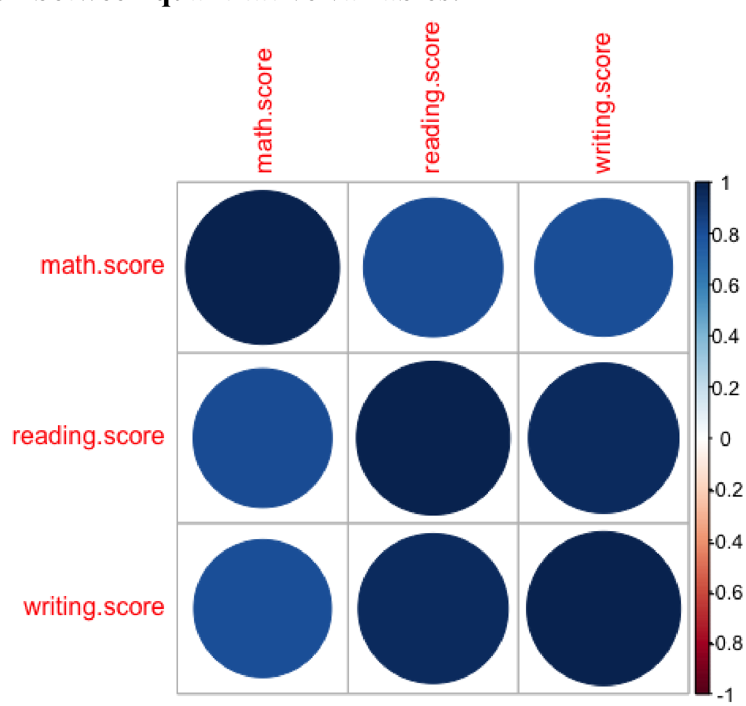
```
   math.score      reading.score    writing.score
 Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
 1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
 Median : 66.00   Median : 70.00   Median : 69.00
 Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
 3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
 Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Distributions of quantitative variables:



Based on the histograms, each variables' distribution approximately represents a normal distribution. This is to be expected due to the Central Limit Theorem.

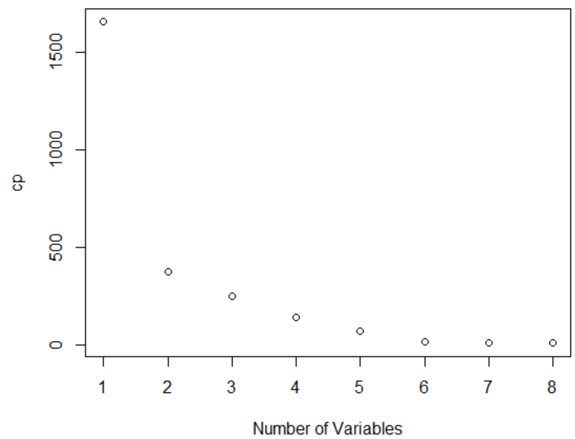
Correlation between quantitative variables:



Each of the test scores variables seem to be highly correlated with each other.

Predicting Math Score:

First we found the best subset of predictors using the “leaps” package in R.



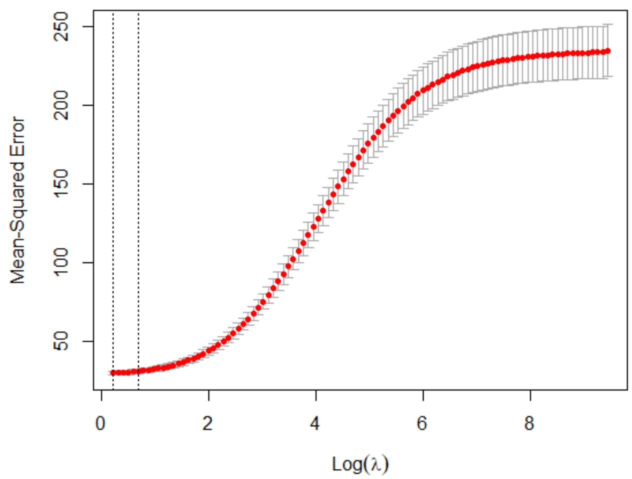
The best subset of predictors (with the lowest cp) is 8 variables, which includes all of the predictors in our data set.

We started by running a <u>linear regression model</u> on this subset of predictors, with math scores as our response variable.

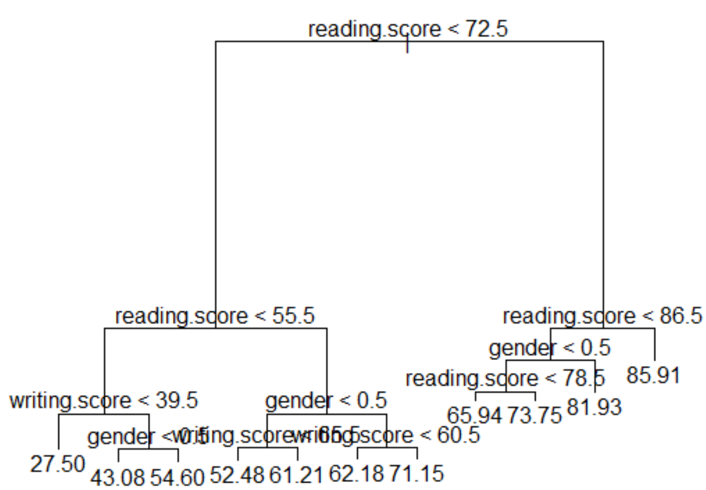
```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              -13.46845    1.77424  -7.591 1.64e-13 ***
gendermale                                13.64231    0.52409  26.030  < 2e-16 ***
race.ethnicitygroup B                      0.84151    1.00004   0.841    0.400    
race.ethnicitygroup C                     -0.27479    0.94184  -0.292    0.771    
race.ethnicitygroup D                      1.06604    0.97922   1.089    0.277    
race.ethnicitygroup E                      4.77527    1.05132   4.542 7.03e-06 ***
parental.level.of.educationbachelor's degree -0.52654  0.88157  -0.597    0.551    
parental.level.of.educationhigh school     0.95131    0.73706   1.291    0.197    
parental.level.of.educationmaster's degree -0.91478   1.08850  -0.840    0.401    
parental.level.of.educationsome college    0.54245    0.71152   0.762    0.446    
parental.level.of.educationsome high school 1.23321   0.76645   1.609    0.108    
lunchstandard                              3.12621    0.51536   6.066 2.64e-09 ***
test.preparation.coursenone                3.78352    0.56352   6.714 5.31e-11 ***
reading.score                              0.24000    0.05946   4.037 6.30e-05 ***
writing.score                              0.74624    0.06241  11.956  < 2e-16 ***
```

The summary above shows that reading and writing scores as well as gender, whether the student belongs to race E, lunch type and whether the student completed test preparation scores are significant predictors of math scores. Our model performed fairly well on the test data with an MSE of 30.95683.

Next we performed Ridge Regression and LASSO. First we used cross validation to find the most optimal lambda which came out to be λ = 1.25.



We calculated the MSE of our models on the test data and Ridge Regression came out to 30.00705 while LASSO came out to 30.46811. Both an improvement from the MSE of our linear regression model.



From the Regression Tree above we can see that three variables are significant. Reading score was the most significant followed by writing and gender. The MSE for the Regression Tree was 57.3639.

We also performed Bagging (Bootstrap Aggregating) and Random Forest and ended up improving our MSE down to 36.96806 and 34.98825 respectively.

```
> importance(random.forest)
                               %IncMSE IncNodePurity
gender                      101.940083     12165.082
race.ethnicity               13.319595      4691.952
parental.level.of.education  -5.549374      2604.227
lunch                        14.396258      4927.554
test.preparation.course       9.835433      1469.678
reading.score                39.538666     48167.785
writing.score                33.494576     38133.004
```

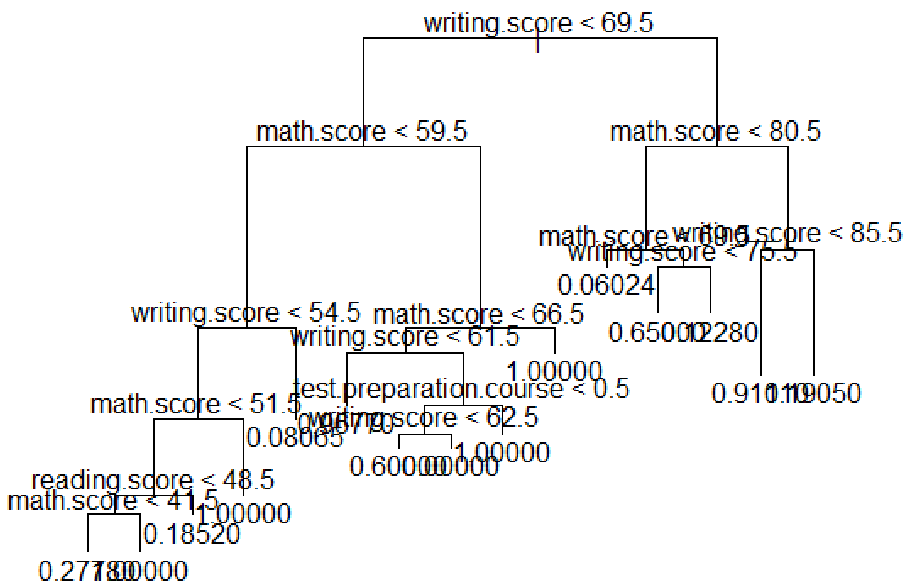
Above is the influence of variables from the random forest method

Predicting Gender:

First we ran a logistic regression including all of the variables.

```
Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                 8.831563   1.634699   5.403 6.57e-08 ***
race.ethnicitygroup B                      -0.916023   0.778746  -1.176  0.23948    
race.ethnicitygroup C                       0.098289   0.704797   0.139  0.88909    
race.ethnicitygroup D                       0.164704   0.758769   0.217  0.82816    
race.ethnicitygroup E                      -2.328379   0.813608  -2.862  0.00421 ** 
parental.level.of.educationbachelor's degree 1.139565   0.744890   1.530  0.12606    
parental.level.of.educationhigh school     -0.366210   0.640422  -0.572  0.56744    
parental.level.of.educationmaster's degree  1.768634   0.922514   1.917  0.05521 .  
parental.level.of.educationsome college     0.554388   0.629769   0.880  0.37869    
parental.level.of.educationsome high school -0.705483   0.684068  -1.031  0.30240    
lunchstandard                              -0.992038   0.458557  -2.163  0.03051 *  
test.preparation.coursenone                -3.782885   0.610375  -6.198 5.73e-10 ***
math.score                                  0.577724   0.061313   9.423  < 2e-16 ***
reading.score                               0.007892   0.051826   0.152  0.87897    
writing.score                              -0.654975   0.082789  -7.911 2.55e-15 ***
```

We decided that predicting a students' gender based on race, parental level of education, and standard/reduced lunch didn't make much sense, and the significance of this model reassured this idea. Our model with all of the predictors had a test error rate of 0.124 while the model with variables removed had a test error rate of 0.132.



From the tree above we can see that the best predictor for gender is writing score and math score. The basic tree error rate is 0.172.

We also performed Bagging (Bootstrap Aggregating) and Random Forest and ended up improving the error rate down to 0.162 for both.

```
> importance(Bagging)
                              %IncMSE IncNodePurity
gender                     133.055685     13516.005
race.ethnicity              14.924731      3128.479
parental.level.of.education -6.619921      1853.215
lunch                       16.081498      1401.176
test.preparation.course     17.682074      1359.841
reading.score               63.916594     66088.504
writing.score               35.265382     27130.870
```

Above is the influence of variables from the random forest method.

Finally we used Support Vector Machines to predict gender. We made models with both a linear and radial kernel and they produced test error rates of 0.116 and 0.174 respectively. These scores were both our best and worst error rates even though there is only a 5.8% difference.

```
Call:
svm(formula = gender ~ ., data = train, kernel = "linear", cost = 10, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10

Number of Support Vectors:  98

 ( 49 49 )


Number of Classes:  2

Levels:
 female male
```

MSE scores for Math Scores **Test error rates for Gender**

Linear Regression: 30.95683	Logistic Regression: 0.124
Ridge: 30.00705	Logistic Regression with removed variables: 0.132
Lasso: 30.46811	Basic Trees: 0.172
Basic Regression Tree: 57.3639	Bagging: 0.162
Bagging: 36.96806	Random Forest: 0.162
Random Forest: 34.98825	SVM Linear: 0.116
	SVM Radial: 0.174

Conclusions:

From our models we can observe that high reading scores predict high math scores, followed by writing scores and gender. Male students performed better on math scores as compared to female students. As for gender, the best predictor ended up being writing score followed up by math scores. Out of all the models we applied, Ridge regression yielded the least MSE for Math Score predictions and Logistic Regression had the least test error rates for Gender prediction. Out of all the predictors Ethnic group C was not significant at all when determining math scores.

Although this was simulated data, we believe it is worthy to do an actual study on these variables as the findings may give insight to how socio economic status and performance in the rest of a students' classes affects their performance in school. It might also be worthy to look into the relationship between gender and test performance in order to gain insight on which subjects one gender outperforms the other, if at all.

https://www.kaggle.com/spscientist/students-performance-in-exams